·论著·

基于机器学习的冠心病风险预测模型构建与比较

岳海涛[1]，何婵婵[1]，成羽攸[1]，张森诚[1]，吴悠[2*]，马晶[1*]

1.518055 广东省深圳市，清华大学医院管理研究院
2.100084 北京市，清华大学医院管理研究院 清华大学医学院
*通信作者：吴悠，助理教授/博士生导师；E-mail：youwu@tsinghua.edu.cn
马晶，教授；E-mail：jingma@sz.tsinghua.edu.cn
岳海涛和何婵婵为共同第一作者

【摘要】 **背景** 冠状动脉粥样硬化性心脏病（Coronary atherosclerotic heart disease，CHD）（以下简称冠心病）是全球重要的死亡原因之一。目前关于冠心病风险评估的研究在逐年增长。然而，在这些研究中常忽略了数据不平衡的问题，而解决该问题对于提高分类算法中识别冠心病风险的准确性至关重要。**目的** 探索冠心病的影响因素，通过使用2种平衡数据的方法，基于5种算法建立冠心病风险相关的预测模型，比较这5种模型对冠心病风险的预测价值。**方法** 基于2021年美国国家行为风险因素监测系统（Behavioral Risk Factor Surveillance System，BRFSS）横断面调查数据筛选出112 606位研究对象的健康相关风险行为、慢性健康状况等24个变量信息，结局指标为自我报告是否患有冠心病并据此分为冠心病组和非冠心病组。通过进行单因素分析和逐步Logistic回归分析探索冠心病发生的影响因素并筛选出纳入预测模型的变量。随机抽取112 606位受访者的10%（共计11 261名），以8：2的比例随机划分为训练与测试的数据集，采用随机过采样（Random oversampling）和合成少数过采样技术（Synthetic Minority Over-sampling Technique，SMOTE）两种过采样（Over-sampling）的方法处理不平衡数据，基于k最邻近算法（K-Nearest Neighbor，KNN）、Logistic回归、支持向量机（Support Vector Machine，SVM）、决策树和XGBoost算法分别建立冠心病预测模型。**结果** 两组年龄、性别、BMI、种族、婚姻状态、教育水平、收入水平、是否被告知患高血压、是否被告知患处于高血压前期、是否被告知患妊娠高血压、现在是否在服用高血压药物、是否被告知患有高血脂、是否被告知患有糖尿病、抽烟情况、过去30 d内是否至少喝过1次酒、是否为重度饮酒者、是否为酗酒者、过去30 d内是否有体育锻炼、心理健康状况以及自我健康评价比较，差异有统计学意义（$P<0.05$）。逐步Logistic回归分析结果显示：年龄、性别、BMI水平、种族、教育水平、收入水平、是否被告知患高血压、是否被告知患处于高血压前期、是否被告知患妊娠高血压、现在是否在服用高血压药物、是否被告知患有高血脂、是否被告知患有糖尿病、抽烟情况、过去30天内是否至少喝过一次酒、是否为重度饮酒者、是否为酗酒者以及自我健康评价为冠心病的影响因素（$P<0.05$）。风险模型构建的分析结果显示：k最邻近算法、Logistic回归、支持向量机、决策树和XGBoost采用合成少数过采样技术处理不平衡数据的总体分类精度分别为59.2%、67.4%、66.2%、69.2%和85.9%；召回率分别为75.2%、71.4%、70.5%、62.9%和34.8%；精确度分别为15.4%、18.2%、17.5%、17.6%和28.7%；F值分别为0.256、0.290、0.280、0.275和0.315；AUC分别为0.80、0.78、0.72、0.72和0.82；采用随机过采样处理不平衡数据的总体分类精度分别为62.5%、68.5%、69.0%、60.2%和70.1%；召回率分别为70.0%、69.5%、71.9%、69.0%和67.6%；精确度分别为15.8%、18.4%、19.1%、14.8%和19.0%；F值分别为0.258、0.291、0.302、0.244和0.297；受试者工作特征曲线下面积分别为0.80、0.77、0.72、0.72和0.83。**结论** 本研究不仅确认了已知冠心病的影响因素，还发现了自我健康评价水平、收入水平和教育水平对冠心病具有潜在影响。在使用2种数据平衡方法后，5种算法的性能显著提高。其中XGBoost模型表现最佳，可作为未来优化冠心病预测模型的参考。此外，鉴于XGBoost模型的优异性能以及逐步Logistic回归的操作便捷和可解释性，推荐在冠心病风险预测模型中，结合使用数据平衡后的XGBoost和逐步Logistic回归分析。

【关键词】 冠心病；机器学习；风险预测模型；Logistic回归；k最邻近算法；支持向量机；决策树；XGBoost

【中图分类号】 R 541.4 **【文献标识码】** A DOI：10.12114/j.issn.1007-9572.2023.0323

Coronary Heart Disease Risk Prediction Model Based On Machine Learning

引用本文：岳海涛，何婵婵，成羽攸，等.基于机器学习的冠心病风险预测模型构建与比较［J］.中国全科医学，2024. DOI：10.12114/j.issn.1007-9572.2023.0323.［Epub ahead of print］［www.chinagp.net］

YUE H T，HE C C，CHENG Y Y，et al.Coronary heart disease risk prediction model based on machine learning［J］. Chinese General Practice，2024.［Epub ahead of print］



YUE Haitao[1]，*HE Chanchan*[1]，*CHENG Yuyou*[1]，*ZHANG Sencheng*[1]，*WU You*[2*]，*MA Jing*[1*]

1.Institute for Hospital Management，Tsinghua University，Shenzhen 518055，China

2.Institute for Hospital Management/ School of Medicine，Tsinghua University，Beijing 100084，China

[*]Corresponding authors：WU You，Assistant Professor，Doctoral supervisor；E-mail：youwu@tsinghua.edu.cn

MA Jing，Professor，Master supervisor；E-mail：jingma@sz.tsinghua.edu.cn

[*]YUE Haitao and HE Chanchan are co-first authors

【Abstract】 **Background** Coronary atherosclerotic heart disease （CHD） is one of the leading causes of mortality worldwide，and research on risk assessment for CHD has been growing annually. However，the issue of data imbalance in these studies is often overlooked，despite its crucial role in enhancing the accuracy of CHD risk identification within classification algorithms. **Objective** To investigate the factors influencing CHD and to establish predictive models for CHD risk using two data balancing methods based on five algorithms，comparing the predictive value of these models for CHD risk. **Methods** Utilizing cross-sectional survey data from the 2021 Behavioral Risk Factor Surveillance System （BRFSS） in the United States，a cohort of 112，606 participants was identified，featuring 24 variables related to risk behaviors and health status，with self-reported coronary heart disease （CHD） as the outcome measure. Factors influencing the incidence of CHD were explored through univariate analysis and stepwise logistic regression to select pertinent variables for inclusion in the predictive model. A random sample comprising 10% of the participants （11，261 individuals） was drawn and then randomly divided into training and testing datasets at an 8：2 ratio. To address data imbalance，two over-sampling techniques were employed：random oversampling and the Synthetic Minority Over-sampling Technique （SMOTE）. Based on these methods，CHD predictive models were constructed using five different algorithms：K-Nearest Neighbors （KNN），Logistic Regression，Support Vector Machine （SVM），Decision Tree，and XGBoost. **Results** Univariate analysis revealed significant differences （P<0.05） between the CHD and non-CHD groups across all input variables except for rental housing and being informed of prediabetic status. Stepwise logistic regression identified age，gender，BMI，ethnicity，education level，income level，being informed of hypertension，being informed of prehypertension，being informed of pregnancy-induced hypertension，current use of antihypertensive medication，being informed of hyperlipidemia，being informed of diabetes，smoking status，alcohol consumption within the last 30 days，heavy drinking status，and self-assessed health as factors influencing CHD. The performance of risk models using SMOTE showed overall classification accuracies of 59.2%，67.4%，66.2%，69.2%，and 85.9%；recall rates of 75.2%，71.4%，70.5%，62.9%，and 34.8%；precision of 15.4%，18.2%，17.5%，17.6%，and 28.7%；F-values of 0.256，0.290，0.280，0.275，and 0.315；and AUC values of 0.80，0.78，0.72，0.72，and 0.82，respectively. Using random oversampling，the models achieved classification accuracies of 62.5%，68.5%，69.0%，60.2%，and 70.1%；recall rates of 70.0%，69.5%，71.9%，69.0%，and 67.6%；precision of 15.8%，18.4%，19.1%，14.8%，and 19.0%；F-values of 0.258，0.291，0.302，0.244，and 0.297；and AUC values of 0.80，0.77，0.72，0.72，and 0.83，respectively. **Conclusion** This study not only confirmed known factors affecting CHD but also identified potential impacts of self-assessed health level，income level，and education level on CHD. The performance of the five algorithms was significantly enhanced after employing two data balancing methods. Among them，the XGBoost model exhibited superior performance and can be referenced for future optimization of CHD prediction models. Additionally，considering the excellent performance of the XGBoost model and the convenience and interpretability of stepwise logistic regression，a combined use of these approaches after data balancing is recommended in CHD risk prediction models.

【Key words】 Coronary Disease；Machine Learning；Risk prediction model；K-nearest neighbor；Support vector machine；Decision tree；Logistic regression；XGBoost

冠状动脉粥样硬化性心脏病又称冠心病，居全球死亡原因之首。2019 年，全球冠心病患者约有 1.97 亿，因冠心病死亡人数约 914 万。中国正面临人口老龄化和心脑血管疾病危险因素增多的双重压力，致使这类疾病的发病率与患病率持续增长[1]。据统计，目前冠心病患者已达 1 139 万例[2]，死亡风险居高不下。在过去三十年间，全球增加的冠心病死亡病例中，38.2% 来自中国[3]。与此同时，高发病率和高死亡率伴随的还有沉重的经济负担：从 1990 年至 2019 年间，全球因冠心病引发的经济负担增加了 82%，年均达到 1.82 亿美元[2]。然而，从健康状态转变为冠心病常通常历时数十年，在此期间有充足的机会采取有效措施进行干预，因此，建立冠心病风险预测模型可以尽早发现患病高危人群，并针对其发病风险进行个性化干预，从而预防冠心病的发生。

目前冠心病的早期风险评估研究在逐年增长，国际上比较成熟的心血管疾病风险预测模型包括弗雷明汉风险评分（FRS）[4]、汇总队列方程[5]以及欧洲冠心病

风险评分系统[6]等，但这些模型开发时间较早，且随着社会经济发展冠心病的影响因素也在发生变化，此外，这些模型主要是基于特定地区的研究，所覆盖的区域较窄。

近年来，随着医疗数据的深入挖掘，越来越多的机器学习算法在冠心病人群中开发、验证[6，7]。然而，大多数研究仅采用了逻辑回归算法[8，9]。此外，许多研究常忽视了数据分布不平衡的问题[10-12]，导致采用整体分类准确性为目标的机器学习算法在训练过程中忽视少数类，使其性能不佳[13]，而采用过采样方法进行样本重构是提高模型性能的关键。此外，一些研究显示部分分类算法在识别风险时准确性较差，且这些研究主要集中于心血管疾病的预测[14]，对于更为细分的冠心病领域研究较少，考虑到冠心病因其因果复杂性，探索其危险因素十分必要。为了解决上述研究中数据集不平衡性问题并探索冠心病的更多潜在影响因素，本研究使用 2021 年美国行为风险因素监测系统（Behavioral Risk Factor Surveillance System，BRFSS）的大规模人群数据，通过采用 2 种过采样方法平衡数据，基于 K 最邻近算法（KNN）、支持向量机（SVM）、决策树、Logistic 回归和 XGBoost，构建冠心病的预测模型并通过混淆矩阵和受试者工作特征（Receiver operating characteristic curve，ROC）曲线确定最优模型。

1 资料与方法

1.1 数据来源

数据集是从 2021 年美国 BRFSS 横断面调查数据中获取[15]。BRFSS 是美国首要的健康相关电话调查系统，主要收集有关美国居民健康相关风险行为、慢性健康状况等[16]。

1.2 研究对象

本研究选取对象为 2021 年美国 BRFSS 的 112 606 位受访者。排除标准：（1）小于 45 岁；（2）本研究的 24 个变量有信息缺失的受访者。采用随机抽样的方式，从 112 606 名受访者中抽取了 10% 的样本，即 11 261 名，以此作为研究的代表性训练和测试集。在样本抽取过程中，设定固定随机种子为 42，以确保抽样的可复制性。

1.3 变量选择

以受访者是否被告知患有冠心病为因变量（输出变量），并以此为根据分为冠心病组和非冠心病组。通过查阅冠心病相关危险因素的相关文献，选择 23 个自变量（输入变量），包括年龄、性别、BMI、种族、婚姻状态、教育水平、收入水平、家里有几个孩子、是否拥有或租用房屋、抽烟情况、过去 30 d 内是否至少喝过 1 次酒、是否为重度饮酒者、是否为酗酒者、自我健康评级、心理健康状况、过去 30 d 内是否有体育锻炼、是否被告知患高血压、是否被告知患妊娠高血压、是否被告知患处于高血压前期、现在是否在服用高血压药物、是否被告知患有高血脂、是否被告知患有糖尿病和是否被告处于糖尿病前期，详见表 1。

1.4 统计学方法

使用 R 4.1.3 和 Python 3.9.13 软件完成所有的数据统计分析，计量资料采用（$\bar{x} \pm s$）表示，计数资料采用相对数表示。通过单因素显著性分析将删除 P>0.01 的变量，此外，采用逐步 Logistic 回归分析确定最终纳入预测模型的变量，每个变量选择一个类别作为对照，并计算其他类别的 *OR* 值和 95% 置信区间（*CI*）。此外，由于受试者中冠心病组占比较低约为 9.35%，属于非平衡数据，为解决数据集不平衡的问题，本研究分别通过随机过采样（Random oversampling）和合成少数过采样技术（Synthetic Minority Over-sampling Technique，SMOTE）处理训练集，其中随机过采样技术是解决数据集类别不平衡问题的一种基本方法，主要通过复制少数类样本以平衡类别分布。而少数类样本合成过采样技术（SMOTE）由 CHAWLA 等[17]于 2002 年提出，该技术通过在位置相近的少数类样本间进行插值生成新样本点，以此来实现数据的平衡，改善模型对少数类别的预测能力。在 2 种采样方法处理后的训练集中利用筛选出的变量选择 k 最邻近算法、支持向量机、决策树、Logistic 回归和 XGBoost 进行建模，各模型训练集和测试集按 8：2 比例，9 009 例样本用于训练，2 252 例样本用于预测。在测试集中采用混淆矩阵和受试者工作特征（ROC）曲线对模型进行评价，所有检验为双侧检验，检验水准 α=0.05。

2 结果

2.1 两组基本特征比较

两组年龄、性别、BMI、种族、婚姻状态、教育水平、收入水平、是否被告知患高血压、是否被告知患处于高血压前期、是否被告知患妊娠高血压、现在是否在服用高血压药物、是否被告知患有高血脂、是否被告知患有糖尿病、抽烟情况、过去 30 d 内是否至少喝过 1 次酒、是否为重度饮酒者、是否为酗酒者、过去 30 d 内是否有体育锻炼、心理健康状况以及自我健康评价比较，差异有统计学意义（*P*<0.05）；两组是否租房以及是否被告知处于糖尿病前期比较，差异无统计学意义（*P*>0.05），见表 2。

2.2 Logistic 回归分析

将单因素分析中 *P*<0.01 的 21 个变量纳入逐步 Logistic 回归模型中进行变量筛选，结果显示，年龄、性别、BMI 水平、种族、教育水平、收入水平、是否被



表 1 变量信息及其赋值
Table 1 Variables information and their assignments

变量	英文问题	中文问题	赋值
年龄	What is your age?	您的年龄是多少?	45~54 岁 =1，55~64 岁 =2，≥ 65 岁 =3
种族	Which one or more of the following would you say is your race?	您属于以下哪一个或多个种族?	白人 =1，黑人 =2，亚洲人 =3，美印第安人 =4，拉丁裔 =5，其他 =6
性别	Are you male or female?	您的性别是男性还是女性?	男 =1，女 =2
婚姻状态	Aer you married?	您已婚吗?	未婚 =1，已婚 =2
教育水平	What is the highest grade or year of school you completed?	您完成的最高学历是?	初中及以下 =1，高中 =2，上过大学或技术学校（没毕业）=3，大学或技术学校毕业 =4
收入水平	Is your annual household income from all sources?	您全家每年从所有来源获得的收入是多少?	<15 000 美元 =1，≥ 15 000 美元且 <25 000 美元 =2，≥ 25 000 美元且 <35 000 美元 =3，≥ 35 000 美元且 <50 000 美元 =4，≥ 50 000 美元且 <100 000 美元 =5，≥ 100 000 美元且 <200 000 美元 =6，≥ 200 000 美元 =7
家里有几个孩子	How many children less than 18 years of age live in your household?	您家里有多少未满 18 岁的孩子?	0 个 =1，1 个 =2，2 个 =3，3 个 =4，≥ 4 个 =5
是否租房	Do you own or rent your home?	您的住房是自有还是租赁?	否 =0，是 =1
是否被告知患高血压	Have you ever been told by a doctor，nurse，or other health professional that you have high blood pressure?	医生、护士或其他健康专业人员是否曾告诉您，您有高血压?	否 =0，是 =1
是否被告知患妊娠高血压	Have you ever been told by a doctor，nurse，or other health professional that you have told only during pregnancy？	医生、护士或其他健康专业人员是否曾告诉您，您只在怀孕期间有高血压?	否 =0，是 =1
是否被告知患处于高血压前期	Have you ever been told by a doctor，nurse，or other health professional that you have told borderline high or prehypertensive or elevated blood pressure?	医生、护士或其他健康专业人员是否曾告诉您，您的血压处于边缘高值或前高血压状态?	否 =0，是 =1
现在是否在服用高血压药物	Are you currently taking prescription medicine for your high blood pressure?	您目前是否正在服用处方药物控制高血压?	否 =0，是 =1
是否被告知患有高血脂	Have you ever been told by a doctor，nurse or other health professional that your cholesterol is high?	医生、护士或其他健康专业人员是否曾告诉您，您的胆固醇水平高?	否 =0，是 =1
是否被告知患有糖尿病	（Ever told）（you had）diabetes?	医生、护士或其他健康专业人员是否曾告诉您，您患有糖尿病?	否 =0，是 =1
是否被告知患处于糖尿病前期	（Ever told）（you had）prediabetes or borderline diabetes?	医生、护士或其他健康专业人员是否曾告诉您，您患有前期糖尿病或血糖边缘升高?	否 =0，是 =1
吸烟情况	Do you now smoke cigarettes every day，some days，or not at all?	您现在是否每天吸烟，偶尔吸烟，还是根本不吸烟?	从不抽烟 =0，已经戒烟 =1，现在偶尔抽烟 =2，现在每天抽烟 =3
过去 30 天内是否至少喝过一次酒	During the past 30 days，on the days when you drank，about how many drinks did you drink on the average?	在过去的 30 天里，您喝酒的日子平均每天喝多少杯?	0 d=0，≥ 1 d=1
是否为重度饮酒者	Heavy drinkers（adult men having more than 14 drinks per week and adult women having more than 7 drinks per week）	成年男性每周饮酒超过 14 杯，成年女性每周饮酒超过 7 杯	否 =0，是 =1
是否为酗酒者	Binge drinkers（males having five or more drinks on one occasion，females having four or more drinks on one occasion）	男性一次性饮酒 5 杯或以上，女性一次性饮酒 4 杯或以上	否 =0，是 =1
过去 30 天内是否有体育锻炼	During the past month，other than your regular job，did you participate in any physical activities or exercises such as running，calisthenics，golf，gardening，or walking for exercise?	在过去的一个月中，除了您的常规工作外，您是否参加过跑步、健身操、高尔夫、园艺或散步等体育活动或锻炼?	否 =0，是 =1
心理健康状况	Now thinking about your mental health，which includes stress，depression，and problems with emotions，for how many days during the past 30 days was your mental health not good?	现在让我们来谈谈您的心理健康，包括压力、抑郁以及情绪问题，在过去 30 d 里，有多少天您的心理健康状况不佳?	非常好（0 d）=1，好（1~7 d）=2，一般（8~14 d）=3，不好（15~21 d）=4，非常不好（22~30 d）=5
自我健康评价	Would you say that in general your health is?	您认为您的总体健康状况如何?	非常不好 =1，不好 =2，一般 =3，好 =4，非常好 =5
BMI	About how much do you weigh without shoes?	您不穿鞋时，您的体重大约是多少?	连续变量
	About how tall are you？	您的身高是多少?	
是否为 CHD	（Ever told）（you had）coronary heart disease?	医生、护士或其他健康专业人员是否曾告诉您，您患有冠心病?	否 =0，是 =1

表 2 冠心病与非冠心病组基本特征比较

Table 2 Sociodemographic characteristics of participants in the coronary and non-coronary groups

项目	非冠心病组（n=112 606）	冠心病组（n=11 261）	χ^2（t）值	P 值
年龄［例（%）］			1 279.863	<0.001
45~54 岁	17 705（17.3）	697（6.7）		
55~64 岁	27 109（26.5）	2 082（20.0）		
≥ 65 岁	57 403（56.2）	7 610（73.3）		
性别［例（%）］			815.090	<0.001
男	42 771（41.8）	5 860（56.4）		
女	59 446（58.2）	4 529（43.6）		
种族［例（%）］			246.436	<0.001
白人	79 561（77.8）	8 619（83.0）		
黑人	8 837（8.6）	685（6.6）		
亚洲人	1 377（1.3）	74（0.7）		
美印第安人	2045（2.0）	206（2.0）		
拉丁裔	7 440（7.3）	439（4.2）		
其他	2 957（2.9）	366（3.5）		
婚姻状态［例（%）］			9.609	0.002
未婚	54 208（53.0）	5 675（54.6）		
已婚	48 009（47.0）	4 714（45.4）		
教育水平［例（%）］			113.105	<0.001
初中及以下	6 990（6.8）	881（8.5）		
高中	31 914（31.2）	3 497（33.7）		
上过大学或技术学校（没毕业）	32 410（31.7）	3 318（31.9）		
大学或技术学校毕业	30 903（30.2）	2 693（25.9）		
收入水平［例（%）］			405.507	<0.001
<15 000 美元	4 078（4.0）	476（4.6）		
≥ 15 000 且 <25 000 美元	5 448（5.3）	788（7.6）		
≥ 25 000 且 <35 000 美元	6 821（6.7）	906（8.7）		
≥ 35 000 且 <50 000 美元	9 694（9.5）	1 277（12.3）		
≥ 50 000 且 <100 000 美元	20 512（20.1）	2 237（21.5）		
≥ 100 000 且 <200 000 美元	24 461（23.9）	2 234（21.5）		
≥ 200 000 美元	31 203（30.5）	2 471（23.8）		
家里有几个孩子［例（%）］			230.841	<0.001
0 个	90 278（88.3）	9 685（93.2）		
1 个	6 624（6.5）	405（3.9）		
2 个	3 432（3.4）	178（1.7）		
3 个	1 269（1.2）	76（0.7）		
≥ 4 个	614（0.6）	45（0.4）		
是否租房［例（%）］			0.047	0.828
否	79 741（78.0）	8 095（77.9）		
是	22 476（22.0）	2 294（22.1）		
是否被告知患高血压［例（%）］			2630.292	<0.001
否	47 417（46.4）	2 096（20.2）		
是	54 800（53.6）	8 293（79.8）		
是否被告知患妊娠高血压［例（%）］			39.430	<0.001
否	100 969（99.6）	10 334(99.8）		
是	1 248（0.4）	55（0.2）		
是否被告知处于高血压前期［例（%）］			6.689	0.010
否	101 801（98.8）	10 364(99.5）		
是	416（1.2）	25（0.5）		
现在是否在服用高血压药物［例（%）］			33.288	<0.001
否	95 954（93.9）	9 899（95.3）		
是	6 263（6.1）	490（4.7）		
是否被告知患有高血脂［例（%）］			2 137.430	<0.001
否	53 563（52.4）	2 971（28.6）		
是	48 654（47.6）	7 418（71.4）		
是否被告知患有糖尿病［例（%）］			1 915.413	<0.001
否	81 820（80.0）	6 387（61.5）		
是	20 397（20.0）	4 002（38.5）		
是否被告知处于糖尿病前期［例（%）］			0.210	0.646
否	99 015（96.9）	10 055(96.8）		
是	3 202（3.1）	334（3.2）		
抽烟情况［例（%）］			813.007	<0.001
从不抽烟	53 198（52.0）	3 999（38.5）		
已经戒烟	33 232（32.5）	4 724（45.5）		
现在偶尔抽烟	3 904（3.8）	412（4.0）		
现在每天抽烟	11 883（11.6）	1 254（12.1）		
过去 30 d 内是否至少喝过 1 次酒［例（%）］			252.923	<0.001
否	59 061（57.8）	6 841（65.8）		
是	43 156（42.2）	3 548（34.2）		
是否为重度饮酒者［例（%）］			68.305	<0.001
否	97 308（95.2）	10 076(97.0）		
是	4 909（4.8）	313（3.0）		
是否为酗酒者［例（%）］			115.244	<0.001
否	93 784（91.7）	9 843（94.7）		
是	8 433（8.3）	546（5.3）		
过去 30 d 内是否有体育锻炼［例（%）］			444.098	<0.001
否	31 781（31.1）	4 282（41.2）		
是	70 436（68.9）	6 107（58.8）		
心理健康状况			285.921	<0.001
非常不好	6 443（6.30）	1 028（9.9）		
不好	902（0.9）	127（1.2）		
一般	5 276（5.2）	701（6.7）		
好	4943（4.8）	553（5.3）		
非常好	84 653（82.8）	7 980（76.8）		
自我健康评价［例（%）］			5459.580	<0.001
非常不好	5 500（5.4）	1 973（19.0）		
不好	17 252（16.9）	3 339（32.1）		
一般	36 093（35.3）	3 334（32.1）		
好	31 637（31.0）	1 469(14.10）		
非常好	11 735（11.5）	274（2.6）		
BMI 水平（$\bar{x} \pm s$，kg/m^2）	29.2157 ± 6.72	30.07 ± 6.78	−12.195[a]	<0.001

注：[a] 表示 t 值。



告知患高血压、是否被告知患处于高血压前期、是否被告知患妊娠高血压、现在是否在服用高血压药物、是否被告知患有高血脂、是否被告知有糖尿病、抽烟情况、过去 30 d 内是否至少喝过 1 次酒、是否为重度饮酒者、是否为酗酒者以及自我健康评价为冠心病的影响因素（P<0.05，见表 3）。

2.3 冠心病预测模型结果分析

本研究采用随机抽样方法，从 112 606 名受访者中选取了 10%（即 11 261 名）的样本，以构建代表性的训练集和测试集。对比总体样本（n=112 606）与随机抽取的样本（n=11 261）在预测模型的 17 个变量上的差异，结果显示差异无统计学意义（P>0.05），详见表 4。此外，按照 8：2 的比例将数据随机抽取样本分为训练集（80%）和测试集（20%），并对训练集样本（n=9 009）与测试集样本（n=2 252）在预测模型的 17 个变量上进行比较，结果显示差异不具有统计学意义（P>0.05），详见表 5。分别利用 5 种算法对原始数据集和平衡后的数据集构建 CHD 预测模型，预测模型的总体分类精度、精确度、召回率、F 值见表 6。在不平衡数据集中使用机器学习方法建模后，测试集中预测模型的召回率、精确度、F 值较低。相比之下，经过数据平衡处理后，机器学习方法建立模型的整体效能提高，尤其是对于阳性样本的分类正确率。在采用 Rand-Oversample 和 SMOTE 过采样方法训练的五种算法中，XGBoost 模型在预测 CHD 方面表现最出色，其测试集的 AUC 值达到 0.83。其次是 KNN 模型，测试集中 AUC 为 0.80。在测试集中，除支持向量机模型与决策树模型在预测 CHD 风险方面表现较差，AUC 仅为 0.72 外，其余几种机器学习算法建立的 CHD 预测模型的效能均较佳。两种不同过采样的方法下的各个模型的测试集工作特征曲线如下图 1 和图 2 所示。

3 讨论

本研究探索了多种冠心病发病的影响因素，通过使用随机过采样和 SMOTE 两种数据平衡方法，建立了基于 5 种不同算法的冠心病风险预测模型，并对其预测价值进行了比较。结果表明，数据平衡显著提升了模型的性能，尤其是 XGBoost 模型在总体分类精度、召回率、精确度、F 值和 AUC 值方面的表现均优于其他模型，显示出其在冠心病风险预测上的强大潜力。

3.1 冠心病发病的影响因素

本研究结果确认了已知的冠心病风险因素，如年龄、性别、高血压、高血脂、糖尿病和吸烟等。但更重要的是发现自我健康评价水平、收入水平和教育水平是冠心病的潜在影响因素。其中，自我健康评价和收入水平对冠心病发病的影响与 OLUSOLA 等[18]和 HEMINGWAY

表 3 Logistic 回归分析结果

Table 3 Results of Logistic regression analysis

变量	B	SE	Wald χ^2 值	P 值	OR（95%CI）
年龄（以 45~54 岁为参照）					
55~64 岁	0.374	0.047	64.120	<0.001	1.454（1.327~1.594）
≥ 65 岁	0.887	0.044	412.301	<0.001	2.428（2.231~2.647）
性别（以男为参照）					
女	-0.554	0.023	600.886	<0.001	0.575（0.550~0.601）
种族（以白人为参照）					
黑人	-0.475	0.044	118.304	<0.001	0.622（0.570~0.677）
亚洲人	-0.581	0.125	21.619	<0.001	0.559（0.434~0.709）
美印第安人	-0.106	0.079	1.807	0.179	0.900（0.769~1.047）
拉丁裔	-0.572	0.055	110.131	<0.001	0.564（0.507~0.627）
其他	0.051	0.061	0.694	0.405	1.052（0.933~1.183）
教育水平（以初中及以下为参照）					
高中	0.039	0.044	0.815	0.367	1.040（0.955~1.134）
上过大学或技术学校（没毕业）	0.11	0.044	6.134	0.013	1.116（0.955~1.219）
大学或技术学校毕业	0.142	0.047	9.216	0.002	1.152（1.052~1.263）
收入水平（以 <15 000 美元为参照）					
≥ 15 000 且 <25 000 美元	0.027	0.065	0.175	0.676	1.028（1.052~1.169）
≥ 25 000 且 <35 000 美元	0.007	0.064	0.012	0.912	1.007（0.889~1.142）
≥ 35 000 且 <50 000 美元	0.045	0.061	0.549	0.459	1.046（0.929~1.142）
≥ 50 000 且 <100 000 美元	-0.048	0.058	0.693	0.405	0.953（0.851~1.142）
≥ 100 000 且 <200 000 美元	-0.134	0.058	5.240	0.022	0.875（0.781~0.982）
≥ 200 000	-0.115	0.059	3.819	0.051	0.891（0.795~1.001）
是否被告知患高血压（以否为参照）					
是	0.751	0.028	729.208	<0.001	2.118（2.006~2.237）
是否被告知处于高血压前期血压（以否为参照）					
是	0.427	0.214	3.986	0.046	1.532（2.006~2.282）
现在是否在服用高血压药物（以否为参照）					
是	-0.482	0.051	90.103	<0.001	0.617（0.558~0.681）
是否被告知患有高血脂（以否为参照）					
是	0.622	0.024	668.066	<0.001	1.863（1.778~1.954）
是否被告知患有糖尿病（以否为参照）					
是	0.327	0.024	180.769	<0.001	1.387（1.322~1.455）
抽烟情况（以从不抽烟为参照）					
已经戒烟	0.342	0.024	197.654	<0.001	1.408（1.342~1.477）
现在偶尔抽烟	0.234	0.058	16.079	<0.001	1.264（1.126~1.416）
现在每天抽烟	0.194	0.038	26.602	<0.001	1.214（1.127~1.306）
过去 30 d 内是否至少喝过一次酒（以否为参照）					
是	-0.096	0.025	14.464	<0.001	0.908（0.864~0.954）
是否为重度饮酒者（以否为参照）					
是	-0.191	0.07	7.405	0.007	0.826（0.719~0.954）
是否为酗酒者（以否为参照）					
是	-0.197	0.056	12.454	<0.001	0.821（0.719~0.915）
自我健康评价（以非常不好为参照）					
不好	-0.546	0.034	255.323	<0.001	0.579（0.719~0.915）
一般	-1.215	0.035	1238.794	<0.001	0.297（0.277~0.318）
好	-1.758	0.041	1840.236	<0.001	0.172（0.159~0.187）
非常好	-2.185	0.07	987.994	<0.001	0.112（0.159~0.129）
BMI 水平（kg/m²）	-0.006	0.002	12.195	<0.001	0.994（0.991~0.997）

表 4 总体样本与随机样本基本特征比较

Table 4 Comparison of general information between the overall sample and random sample

项目	总样本（n=112 606）	随机样本（n=11 261）	χ^2（t）值	P 值
年龄［例（%）］			0.063	0.969
45~54 岁	65 013（57.7）	6 504（57.8）		
55~64 岁	18 402（16.3）	1 848（16.4）		
≥ 65 岁	29 191（25.9）	2 909（25.9）		
性别［例（%）］			0.546	0.460
男	48 631（43.2）	4 904（43.5）		
女	63 975（56.8）	6 357（56.5）		
种族［例（%）］			4.341	0.501
白人	88 180（78.3）	8 783（78.0）		
黑人	9 522（8.5）	942（8.4）		
亚洲人	1 451（1.3）	168（1.5）		
美印第安人	2 251（2.0）	238（2.1）		
拉丁裔	7 879（7.0）	789（7.0）		
其他	3 323（3.0）	341（3.0）		
教育水平［例（%）］			1.593	0.661
初中及以下	7 871（7.0）	798（7.1）		
高中	35 411（31.4）	3 522（31.3）		
上过大学或技术学校（没毕业）	35 728（31.7）	3 628（31.2）		
大学或技术学校毕业	33 596（29.8）	3 313（29.4）		
收入水平［例（%）］			5.847	0.440
<15 000 美元	4 554（4.0）	479（4.3）		
≥ 15 000 且<25 000 美元	6 236（5.5）	626（5.6）		
≥ 25 000 且<35 000 美元	7 727（6.9）	758（6.9）		
≥ 35 000 且<50 000 美元	10 971（9.7）	1 097（9.7）		
≥ 50 000 且<100 000 美元	22 749（20.2）	2 211（19.6）		
≥ 100 000 且<200 000 美元	26 695（23.7）	2 756（24.5）		
≥ 200 000 美元	33 674（29.9）	3 334（29.6）		
是否被告知患高血压［例（%）］			0.304	0.581
否	49 513（44.0）	4 921（43.7）		
是	63 093（56.0）	6 340（56.3）		
是否被告知患处于高血压前期［例（%）］			0.492	0.483
否	112 165（99.6）	11 212（99.6）		
是	441（0.4）	49（0.4）		
现在是否在服用高血压药物［例（%）］			0.046	0.830
否	105 853（94.0）	10 580（94.0）		
是	6 753（6.0）	681（6.0）		
是否被告知患有高血脂［例（%）］			0.077	0.782
否	56 534（50.2）	5 669（50.3）		
是	56 072（49.8）	5 592（49.7）		
是否被告知患有糖尿病［例（%）］			0.724	0.395
否	88 207（78.3）	8 860（78.7）		
是	24 399（21.7）	2 401（21.3）		
抽烟情况［例（%）］			6.211	0.102
从不抽烟	57 197（50.8）	5 745（51.0）		
已经戒烟	37 956（33.7）	3 856（34.2）		
现在偶尔抽烟	4 316（3.8）	433（3.8）		
现在每天抽烟	13 137（11.7）	1 227（10.9）		
过去 30 天内是否至少喝过一次酒［例（%）］			0.051	0.821
否	65 902（58.5）	6 578（58.4）		
是	46 704（41.5）	4 683（41.6）		
是否为重度饮酒者［例（%）］			0.019	0.891
否	107 384（95.4）	10 742（95.4）		
是	5 222（4.6）	519（4.6）		
是否为酗酒者［例（%）］			0.072	0.789
否	103 627（92.0）	10 355（92.0）		
是	8 979（8.0）	906（8.0）		
自我健康评价［例（%）］			1.938	0.747
非常不好	7 473（6.6）	749（6.7）		
不好	20 591（18.3）	2 044（18.2）		
一般	39 427（35.0）	3 917（34.8）		
好	33 106（29.4）	3 376（30.0）		
非常好	12 009（10.7）	1 175（10.4）		
BMI 水平（$\bar{x} \pm s$，kg/m^2）	29.29 ± 6.73	29.33 ± 6.74	−0.673[a]	0.779

注：[a] 表示 t 值。

等[19]的研究结果相同。尽管自我健康评价是主观的评价指标，但在流行病学和健康经济学研究中，此指标已被证实与死亡率、住院率及慢性病发病率等客观健康指标密切相关。例如，DESALVO 的研究发现，自我健康评价与死亡风险之间存在显著关系，即使在控制了其他健康指标后，这种关系仍然存在[20]。MAVADDAT 的系统综述中也发现，在既往有和没有心血管疾病的人群中，自评健康状况不佳都与心血管死亡率有关[21]。因此，自我健康评价在个体冠心病风险的预测中具有不容忽视的价值。然而，考虑到自我健康评价可能受到个人主观感受的影响，存在一定程度的主观偏差，在应用这一指标时应保持谨慎。教育水平与 TAAVI 等[22]的研究结果相反，这可能是因为教育水平高的人虽然健康意识更高，但其职业压力和不健康的饮食习惯更多，这一定程度上也会增加冠心病的风险。此外，本研究发现患有妊娠高血压或糖尿病前期并不能增加患冠心病的风险。这与当前许多研究结果不一致，如妊娠高血压[23]以及糖尿病前期[24]与冠心病有关的研究结果。但是，本研

表 5 训练集样本与测试集样本基本特征比较

Table 5 Comparison of general information between the training set and test set samples

项目	训练集（*n*=9 009）	测试集（*n*=2 252）	χ^2（*t*）值	*P* 值
年龄［例（%）］			0.901	0.637
45~54 岁	5 184（57.5）	1 320（58.6）		
55~64 岁	1 489（16.5）	359（15.9）		
≥ 65 岁	1 489（25.9）	359（25.4）		
性别［例（%）］			0.063	0.802
男	5 091（56.5）	1 266（56.2）		
女	3 918（43.5）	986（43.8）		
种族［例（%）］			4.190	0.522
白人	7 032（78.1）	1 751（77.8）		
黑人	741（8.2）	201（8.9）		
亚洲人	136（1.5）	32（1.4）		
美印第安人	200（2.2）	38（1.7）		
拉丁裔	624（6.9）	165（7.3）		
其他	276（3.1）	65（2.9）		
教育水平［例（%）］			0.900	0.993
初中及以下	641（7.1）	157（7.0）		
高中	2 819（31.3）	703（31.2）		
上过大学或技术学校（没毕业）	2 898（32.2）	730（32.4）		
大学或技术学校毕业	2 651（29.4）	662（29.4）		
收入水平［例（%）］			6.916	0.329
<15 000 美元	389（4.3）	90（4.0）		
≥ 15 000 且 <25 000 美元	522（5.8）	104（4.6）		
≥ 25 000 且 <35 000 美元	598（6.6）	160（7.1）		
≥ 35 000 且 <50 000 美元	862（9.6）	235（10.4）		
≥ 50 000 且 <100 000 美元	1 767（19.6）	444（19.7）		
≥ 100 000 且 <200 000 美元	2 206（24.5）	550（24.4）		
≥ 200 000 美元	2 665（29.6）	669（29.7）		
是否被告知患高血压［例（%）］			0.187	0.665
否	3 946（43.8）	975（43.3）		
是	5 063（56.2）	1 277（56.7）		
是否被告知患处于高血压前期［例（%）］			0.001	0.969
否	8 876（98.5）	2 219（98.5）		
是	133（1.5）	33（1.5）		
现在是否在服用高血压药物［例（%）］			1.223	0.269
否	8 453（93.8）	2 127（94.4）		
是	556（6.2）	125（5.6）		
是否被告知患有高血脂［例（%）］			0.393	0.531
否	4 522（50.2）	1 147（50.9）		
是	4 487（49.8）	1 105（49.1）		
是否被告知患有糖尿病［例（%）］			0.259	0.611
否	7 097（78.8）	1 763（78.3）		
是	1 912（21.2）	489（21.7）		
抽烟情况［例（%）］			6.129	0.106
从不抽烟	4 549（50.5）	1 196（53.1）		
已经戒烟	3 131（34.8）	725（32.2）		
现在偶尔抽烟	343（3.8）	90（4.0）		
现在每天抽烟	986（10.9）	241（10.7）		
过去 30 d 内是否至少喝过 1 次酒［例（%）］			0.046	0.830
否	5 267（58.5）	1 311（58.2）		
是	3 742（41.5）	941（41.8）		
是否为重度饮酒者［例（%）］			0.062	0.804
否	8 596（95.4）	2 146（95.3）		
是	413（4.6）	106（4.7）		
是否为酗酒者［例（%）］			0.011	0.918
否	8 283（91.9）	2 072（92.0）		
是	726（8.1）	180（8.0）		
自我健康评价［例（%）］			3.294	0.510
非常不好	616（6.8）	133（5.9）		
不好	1 626（18.0）	418（18.6）		
一般	3 126（34.7）	791（35.1）		
好	2 710（30.1）	666（29.6）		
非常好	931（10.3）	244（10.8）		
BMI 水平（$\bar{x} \pm s$，kg/m^2）	29.35 ± 6.78	29.28 ± 6.49	0.428[a]	0.668

注：[a] 表示 *t* 值。

表 6 冠心病风险预测模型的预测效能指标

Table 6 Indicators of predictive efficacy for the model predicting the risk of coronary heart disease

模型	总体分类精度	召回率	精确度	*F* 值	AUC
SMOTE					
KNN	0.592	0.752	0.154	0.256	0.800
逻辑回归	0.674	0.714	0.182	0.290	0.770
支持向量机	0.662	0.705	0.175	0.280	0.720
决策树	0.692	0.629	0.176	0.275	0.720
XGBoost	0.859	0.348	0.287	0.315	0.830
Random over-sampling					
KNN	0.625	0.700	0.158	0.258	0.800
逻辑回归	0.685	0.695	0.184	0.291	0.780
支持向量机	0.69	0.719	0.191	0.302	0.720
决策树	0.602	0.690	0.148	0.244	0.720
XGBoost	0.701	0.676	0.190	0.297	0.820
Unbalanced					
KNN	0.907	0	NaN	NaN	NaN
逻辑回归	0.907	0	NaN	NaN	NaN
支持向量机	0.907	0	NaN	NaN	NaN
决策树	0.907	0	NaN	NaN	NaN
XGBoost	0.908	0.029	0.600	0.055	NaN

注：AUC= 受试者工作特征曲线下面积。

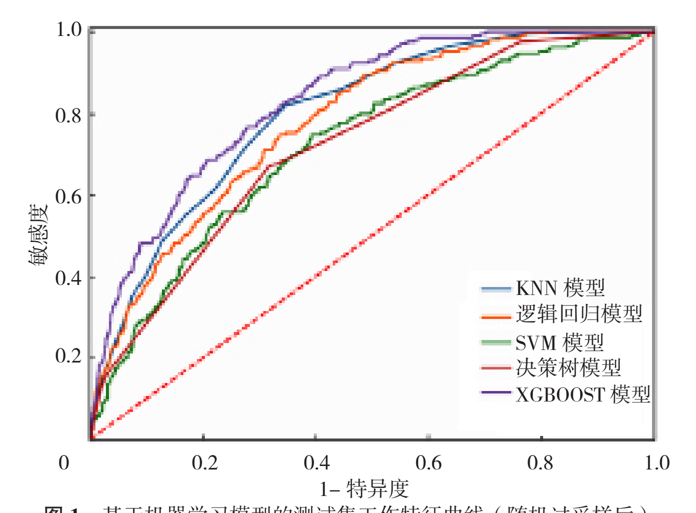


图 1 基于机器学习模型的测试集工作特征曲线（随机过采样后）
Figure 1 Working feature curve of the test set using a machine learning model with random oversampling

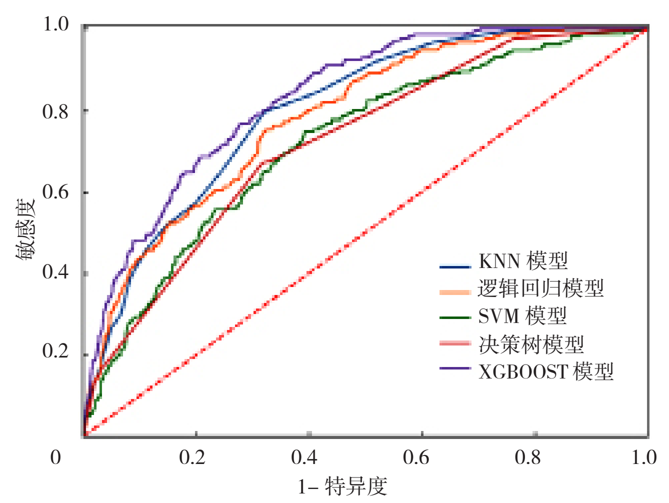


图 2 基于机器学习模型的测试集工作特征曲线（合成少数过采样后）
Figure 2 Working feature curve of the test set using a machine learning model with SMOTE oversampling

究结果与 OHSAK 等[25]关于高血压前期与心脑血管疾病的关系研究结果一致。同时，当前服用高血压药物能够降低冠心病的发病风险，这与 CORRAO 等[26]的研究结果一致。无论是已经戒烟、现在偶尔抽烟还是现在每天抽烟都能够显著增加患冠心病的风险，并且已经戒烟的人群中患冠心病的风险更高，与相关研究一致[27]。关于喝酒对冠心病的影响，本研究发现过去 30 d 内至少喝过一次、重度饮酒以及酗酒都能显著降低冠心病的发病风险，目前比较的成熟的研究表明平均饮酒量对冠心病的影响呈 J 型，轻度至中度饮酒者的冠心病发病风险比戒酒者低，但重度饮酒者中的冠心病风险最高[28]，此外，ZHAO 等[29]的研究也证实与不饮酒者相比，轻度至中度饮酒者能够显著降低心血管疾病的死亡率。

3.2 冠心病风险预测模型

本研究结果表明在经过不平衡数据的处理后，5 种算法在准确率和稳定性方面都有了显著提高，其中 XGBoost 模型对 CHD 的预测效能最佳，这与 ZHANG 等[30]在冠状动脉疾病风险预测模型构建中得到的结论一致，而支持向量机模型与决策树模型在预测 CHD 风险方面表现较差。因为相比其他四种机器学习方法，XGBoost 采用了更为优秀的树结构优化方法，例如叶子节点权重缩减等[31]。此外，在每次训练模型时，XGBoost 还会对模型进行正则化，以避免过拟合，并且使用了一种经过优化的目标函数，这能够更有效地控制决策树的生成，从而提高模型的精度[32]。此外，XGBoost 在训练过程中还能够对特征进行子采样并且还支持并行计算，能够利用多核中央处理机和分布式环境加速模型训练[33]。在模型的总体分类精度表现方面 k 最邻近算法得分较低，这与 HASSAN 等[34]的研究结果一致，但其简单性和易于实现可能在某些情况下具有优势[35]，虽然决策树算法的准确率略低，但其是一种成熟的算法，在 KIM 等[36]的研究发现决策树算法在冠心病风险预测表现较好，本研究结果与其不一致的原因可能是决策树算法对输入数据的微小变化很敏感，这可能导致树结构和预测结果发生很大变化并且进一步降低 CHD 风险预测模型的稳定性，从而导致准确性降低[37]。支持向量机算法在精准度、召回率和 F1 评分方面表现良好，GARAVAND 等[38]在使用多种机器学习方法构建冠心病的诊断模型中也得到相同的结论，可见使用支持向量机算法构建的预测模型可以有效降低疾病诊断的假阳性率。逻辑回归在精度和稳定性方面表现出良好的性能，所以在可成为大规模人群预测模型研究的最佳选择。虽然 XGBoost 是五种算法中冠心病风险预测的最佳算法，但算法的选择将取决于预测任务的具体要求，其较低的召回率表示在实际使用的过程中可能无法筛查出阳性。一般来说，算法的性能受到数据集的大小和质量以及使用的预测变量的数量的影响[38]。本研究强调了不同算法在冠心病风险预测中的优势和劣势，未来计划扩展研究，包括更多的算法和更大的数据集，以更好地评估这些算法的性能。

3.3 研究优点和局限性

本研究的一个显著特点是采用了先进的数据平衡技术，并探索了多种机器学习算法在处理冠心病数据时的效能。优化后的 XGBoost 模型不仅性能出色，且考虑到其与逐步 Logistic 回归分析的综合应用，为临床实践提供了更高的可行性和准确性。此外本次研究有模型输入变量都来自于自我报告，所以具有可及性极高、数据信噪比较低以及贴近实际应用场景等优势。然而，本研究的方法也存在一定的局限性。首先由于机器学习模型的构建基于逻辑回归筛选后的变量，可能忽略了一些未被

初步选中但实际上对疾病预测有潜在价值的变量。其次本文的研究数据来源自美国人群自我报告的横断面BRFSS数据库，数据的可获取性和实际应用场景的贴近性是其优势所在。然而，此类数据可能受到回忆偏差的影响，且横断面研究的设计限制了对因果关系的探讨，并存在结局变量影响输入变量的可能。此外，研究中仅考察了有限的预测模型，并未包括已验证有效的其他模型如人工神经网络（Artificial Neural Network，ANN）、随机森林和朴素贝叶斯。尽管有研究显示CatBoost和LightGBM等梯度提升模型在风险预测上有优秀表现，但本研究未能涉及[39-40]。

4 小结

在处理不平衡数据后，KNN、SVM、决策树、逻辑回归和XGBoost这五种机器学习算法在预测冠心病风险方面的性能可以显著提高，XGBoost是这五种算法中表现最佳的。但是，算法的选择将取决于预测任务的具体要求，以及考虑预测准确性、计算能力、可解释性和运行效率等方面[41]。此外，虽然本研究所采用的方法论和策略在其他人群中具有普适性，但特定的风险预测结果并不适宜直接应用于非美国地区。未来的研究可以通过纳入更广泛的预测模型和多元化的数据来源，来提高模型的普适性和预测的精确度。

作者贡献：岳海涛和何婵婵负责数据分析及解读、文章撰写、文章修改；成羽攸和张森诚协助数据分析及文章撰写。马晶和吴悠负责研究设计、文章的质量控制与审查，监督管理。

本文无利益冲突。

参考文献

[1] 马丽媛，王增武，樊静，等.《中国心血管健康与疾病报告2021》要点解读［J］. 中国全科医学，2022，25（27）：3331-3346. DOI：10.12114/j.issn.1007-9572.2022.0506.

[2] SAFIRI S，KARAMZAD N，SINGH K，et al. Burden of ischemicheart disease and its attributable risk factors in 204 countries and territories，1990-2019［J］. Eur J Prev Cardiol，2022，29（2）：420-431. DOI：10.1093/eurjpc/zwab213.

[3] DAI H J，MUCH A A，MAOR E，et al. Global，regional，and national burden of ischaemic heart disease and its attributable risk factors，1990-2017：results from the Global Burden of Disease Study 2017［J］. Eur Heart J Qual Care Clin Outcomes，2022，8（1）：50-60. DOI：10.1093/ehjqcco/qcaa076.

[4] D'AGOSTINO R B Sr，VASAN R S，PENCINA M J，et al. General cardiovascular risk profile for use in primary care：the Framingham Heart Study［J］. Circulation，2008，117（6）：743-753. DOI：10.1161/CIRCULATIONAHA.107.699579.

[5] ANDRUS B，LACAILLE D. 2013 ACC/AHA guideline on the assessment of cardiovascular risk［J］. J Am Coll Cardiol，2014，63（25 Pt A）：2886. DOI：10.1016/j.jacc.2014.02.606.

[6] CUENDE J I，CUENDE N，CALAVERAS-LAGARTOS J. How to calculate vascular age with the SCORE project scales：a new method of cardiovascular risk evaluation［J］. Eur Heart J，2010，31（19）：2351-2358. DOI：10.1093/eurheartj/ehq205.

[7] ZHANG L Y，NIU M M，ZHANG H Y，et al. Nonlaboratory-based risk assessment model for coronary heart disease screening：model development and validation［J］. Int J Med Inform，2022，162：104746. DOI：10.1016/j.ijmedinf.2022.104746.

[8] AMBRISH G，GANESH B，GANESH A，et al. Logistic regression technique for prediction of cardiovascular disease［J］. Glob Transitions Proc，2022，3（1）：127-130. DOI：10.1016/j.gltp.2022.04.008.

[9] 马婧怡，刘相佟，吕世云，等. 北京市成年人冠心病七年发病风险评估与预测模型［J］. 心肺血管病杂志，2022，41（1）：25-30，50. DOI：10.3969/j.issn.1007-5062.2022.01.006.

[10] 李婕，向菲. 冠心病风险因素识别及其预测模型构建［J］. 中华医学图书情报杂志，2020，29（6）：7-13. DOI：10.3969/j.issn.1671-3982.2020.06.002.

[11] 王晓丽，施天行，彭德荣，等. 两种机器学习算法构建老年冠心病患病风险评估模型的效能比较研究［J］. 中华全科医学，2021，19（4）：523-527. DOI：10.16766/j.cnki.issn.1674-4152.001852.

[12] 谭志军，梁英，石福艳，等. 基于企业健康管理的劳动力人群冠心病风险评估模型研究［J］. 现代预防医学，2017，44（12）：2192-2195，2210.

[13] ALI H，MOHD SALLEH M N，SAEDUDIN R，et al. Imbalance class problems in data mining：a review［J］. Indones J Electr Eng Comput Sci，2019，14（3）：1552. DOI：10.11591/ijeecs.v14.i3.pp1552-1563.

[14] 齐俊锋，韩胜红，李俊琳，等. 心血管病危险因素的风险特征分析及疾病预测模型研究［J］. 现代预防医学，2022，49（18）：3283-3287. DOI：10.20043/j.cnki.MPM.202111359.

[15] Delaware Department of Health and Social Services，Division of Public Health，Delaware Behavioral Risk Factor Survey （BRFS）［M］. 2021.

[16] The Behavioral Risk Factor Surveillance System［EB/OL］.（2022-12-2）［2023-04-30］.https：//www.cdc.gov/brfss/index.html.

[17] CHAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：synthetic minority over-sampling technique［J］. Jair，2002，16：321-357. DOI：10.1613/jair.953.

[18] ORIMOLOYE O A，MIRBOLOUK M，IFTEKHAR UDDIN S M I，et al. Association between self-rated health，coronary artery calcium scores，and atherosclerotic cardiovascular disease risk：the multi-ethnic study of atherosclerosis （MESA）［J］. JAMA Netw Open，2019，2（2）：e188023. DOI：10.1001/jamanetworkopen.2018.8023.

[19] HEMINGWAY A. Determinants of coronary heart disease risk for women on a low income：literature review［J］. J Adv Nurs，2007，60（4）：359-367. DOI：10.1111/j.1365-2648.2007.04418.x.

[20] DESALVO K B，BLOSER N，REYNOLDS K，et al. Mortality prediction with a single general self-rated health question. A meta-

analysis [J]. J Gen Intern Med, 2006, 21 (3): 267-275. DOI: 10.1111/j.1525-1497.2005.00291.x.

[21] MAVADDAT N, PARKER R A, SANDERSON S, et al. Relationship of self-rated health with fatal and non-fatal outcomes in cardiovascular disease: a systematic review and meta-analysis [J]. PLoS One, 2014, 9 (7): e103509. DOI: 10.1371/journal.pone.0103509.

[22] TILLMANN T, VAUCHER J, OKBAY A, et al. Education and coronary heart disease: Mendelian randomisation study [J]. BMJ, 2017, 358: j3542. DOI: 10.1136/bmj.j3542.

[23] CANOY D, CAIRNS B J, BALKWILL A, et al. Hypertension in pregnancy and risk of coronary heart disease and stroke: a prospective study in a large UK cohort [J]. Int J Cardiol, 2016, 222: 1012-1018. DOI: 10.1016/j.ijcard.2016.07.170.

[24] HUANG Y L, CAI X Y, MAI W Y, et al. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis [J]. BMJ, 2016, 355: i5953. DOI: 10.1136/bmj.i5953.

[25] HOZAWA A, KURIYAMA S, KAKIZAKI M, et al. Attributable risk fraction of prehypertension on cardiovascular disease mortality in the Japanese population: the Ohsaki Study [J]. Am J Hypertens, 2009, 22 (3): 267-272. DOI: 10.1038/ajh.2008.335.

[26] CORRAO G, NICOTRA F, PARODI A, et al. Cardiovascular protection by initial and subsequent combination of antihypertensive drugs in daily life practice [J]. Hypertension, 2011, 58 (4): 566-572. DOI: 10.1161/HYPERTENSIONAHA.111.177592.

[27] BOUABDALLAOUI N, MESSAS N, GREENLAW N, et al. Impact of smoking on cardiovascular outcomes in patients with stable coronary artery disease [J]. Eur J Prev Cardiol, 2021, 28 (13): 1460-1466. DOI: 10.1177/2047487320918728.

[28] MATSUMOTO-YAMAUCHI H, KONDO K, MIURA K, et al. Relationships of alcohol consumption with coronary risk factors and macro- and micro-nutrient intake in Japanese people: the INTERLIPID study [J]. J Nutr Sci Vitaminol, 2021, 67 (1): 28-38. DOI: 10.3177/jnsv.67.28.

[29] ZHAO J H, STOCKWELL T, ROEMER A, et al. Alcohol consumption and mortality from coronary heart disease: an updated meta-analysis of cohort studies [J]. J Stud Alcohol Drugs, 2017, 78 (3): 375-386. DOI: 10.15288/jsad.2017.78.375.

[30] ZHANG S S, YUAN Y Y, YAO Z H, et al. Improvement of the performance of models for predicting coronary artery disease based on XGBoost algorithm and feature processing technology [J]. Electronics, 2022, 11 (3): 315. DOI: 10.3390/electronics11030315.

[31] LIU J L, WU J F, LIU S R, et al. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model [J]. PLoS One, 2021, 16 (2): e0246306. DOI: 10.1371/journal.pone.0246306.

[32] ZHENG H T, YUAN J B, CHEN L. Short-term load forecasting using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation [J]. Energies, 2017, 10 (8): 1168. DOI: 10.3390/en10081168. DHALIWAL S, NAHID A A, ABBAS R. Effective intrusion detection system using XGBoost [J]. Information, 2018, 9 (7): 149. DOI: 10.3390/info9070149.

[33] HASSAN C A U, IQBAL J, IRFAN R, et al. Effectively predicting the presence of coronary heart disease using machine learning classifiers [J]. Sensors, 2022, 22 (19): 7227. DOI: 10.3390/s22197227.

[34] PAPANIKOLAOU M, EVANGELIDIS G, OUGIAROGLOU S. Dynamic k determination in k-NN classifier: a literature review [C] //2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA). Chania Crete, Greece. IEEE, 2021: 1-8. DOI: 10.1109/IISA52424.2021.9555525.

[35] KIM J, LEE J, LEE Y. Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree [J]. Healthc Inform Res, 2015, 21 (3): 167-174. DOI: 10.4258/hir.2015.21.3.167.

[36] TEAM C.Decision Tree [EB/OL]. (2022-11-29) [2023-02-12]. https: //corporatefinanceinstitute.com/resources/data-science/decision-tree/

[37] GARAVAND A, SALEHNASAB C, BEHMANESH A, et al. Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms [J]. J Healthc Eng, 2022, 2022: 5359540. DOI: 10.1155/2022/5359540.

[38] JAISWAL S, GUPTA P. Ensemble approach: XGBoost, CATBoost, and LightGBM for diabetes mellitus risk prediction [C] //2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA). Gunupur, India. IEEE, 2022: 1-6. DOI: 10.1109/ICCSEA54677.2022.9936130.

[39] 白皙，罗云云，周智博，等. 基于机器学习算法的大于胎龄儿风险预测模型 [J]. 中华流行病学杂志，2021，42 (12): 2143-2148. DOI: 10.3760/cma.j.cn112338-20210824-00677.

[40] 马倩倩，孙东旭，石金铭，等. 基于支持向量机与 XGboost 的成年人群肿瘤患病风险预测研究 [J]. 中国全科医学，2020，23 (12): 1486-1491. DOI: 10.12114/j.issn.1007-9572.2020.00.066.

（收稿日期：2023-09-16；修回日期：2023-12-11）

（本文编辑：崔莎）